NATIONAL CENTER FOR THE MULTISCALE ANALYSIS OF GENOMIC AND CELLULAR NETWORKS

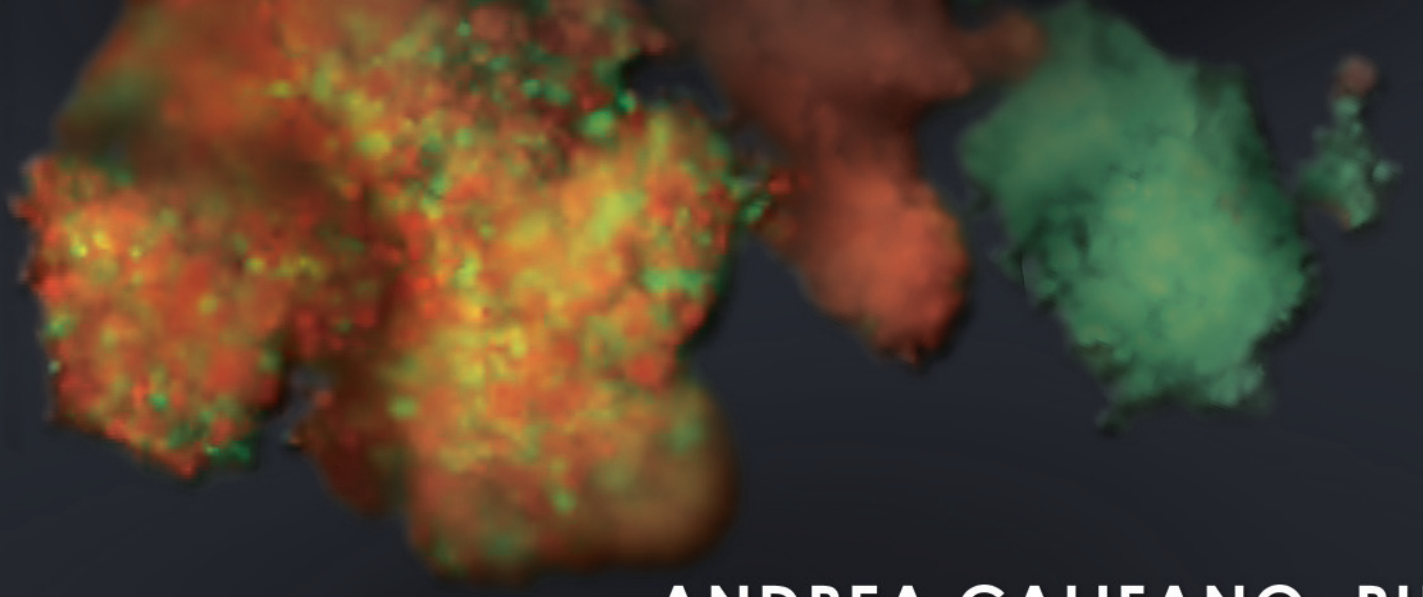
WINTER 2008 ISSUE NO.1 VOL. 1

MAGNet NEWSLETTER

UNDERSTANDING CADHERIN SPECIFICITY IN THE DEVELOPMENT OF MULTICELLULAR STRUCTURES

A COMBINED EXPERIMENTAL AND COMPUTATIONAL STUDY

LAWRENCE SHAPIRO, PHD & BARRY HONIG, PHD



ANDREA CALIFANO, PHD
RICCARDO DALLA-FAVERA, MD

MAPPING THE TRANSCRIPTION FACTOR MODULATOR REPERTOIRE IN HUMAN B LYMPHOCYTES

DAN GALLAHAN, PHD

NATIONAL CENTERS FOR BIOMEDICAL COMPUTING AND RELATED ACTIVITIES AT NIH

WINTER 2008
ISSUE NO.1, VOL 1

MAGNet Newsletter
http://magnet.c2b2.columbia.edu

MAGNet NEWSLETTER

FEATURES

SECTIONS

INTRODUCTION

Welcome to the first edition of the Columbia MAGNet Center newsletter. The Center for the Multiscale Analysis of Genetic and cellular Networks is one of the seven National Center for Biomedical Computing (NCBC) funded by the NIH Roadmap. Our main goal, in collaboration with the other NCBCs, to create the very fabric of a national biomedical computing infrastructure, providing innovative computational methodology and tools to help molecular biology move into the 21st century. As Physics, Chemistry, and Economics, just to name a few disciplines, have evolved from a completely empirical model to one where the interplay between theory and experimentation is much more balanced, we envisage the future of Biology and Medicine to be eventually located at the boundary between the computational and the experimental sciences. This emerging integrative model is intimately reflected in the structure of the NCBCs' scientific programs and educational initiatives. An important goal, for instance, is to create a new breed of researchers trained in both experimental and computational biology. These will be complemented by a vast and integrated array of tools that will support their research. In the opening article of this first issue, Dr. Daniel Gallahan, MAGNet program director, supports this view by presenting a unique NIH perspective of why computation is not just important, but is in fact essential to biomedical research. Dr. Gallahan also summarizes the rationale for the creation of the National Centers for Biomedical Computing.

While the NCBCs are highly integrated and complementary, each one maintains its own identity and is markedly distinct from the others. This is one of the significant strengths of this program. It allows centers to collaborate, rather than compete with each other, while covering an extraordinary range of inter-related activities at the intersection of computation, Biology, and Medicine. MAGNet, specifically, addresses the increasingly important issue of how one may systematically map the molecular interactions underlying inter- and intra-cellular processes within different organisms and cellular phenotypes, using a variety of clues including structural and functional ones. Furthermore, it investigates how these interaction models can be leveraged to dissect normal and disease related processes, laying the path to new biomedical knowledge and discovery.

To illustrate these broad goals, we feature two articles that span the complete spectrum of activities within MAGNet, providing a glimpse of the true multiscale nature of the problems we are tackling. In the first article, Drs. Califano and Dalla Favera discuss how high-throughput biological data and information theory can go hand in hand to help identify key proteins that change the cell regulatory logic at the post-translational level in human B cells. These proteins, including those in signaling and proteolytic pathways, are involved in lymphomagenesis and tumor progression. Additionally, many of them can be targeted by drugs thus allowing a more rational approach to the development of cancer therapies. In the second article, Drs. Shapiro and Honig discuss how minute changes in affinity between different cadherins may produce macroscopic effects by providing cells with exquisitely specific adhesion properties, affecting normal and pathologic processes. The ultimate goal of this project is to understand the molecular basis of specificity, affecting an extraordinarily large range of cellular processes.

Finally, a variety of small featured news articles will provide samples of activities within MAGNet as well as key pointers on how to access the MAGNet tools and infrastructure through the geWorkbench platform for integrative biomedical research.

- Andrea Califano, Ph.D.

NATIONAL CENTERS FOR BIOMEDICAL COMPUTING AND RELATED ACTIVITIES AT NIH

DAN GALLAHAN, PHD

DEPUTY DIRECTOR, DIVISION OF CANCER BIOLOGY, NATIONAL CANCER INSTITUTE

Over the past several years there has been a revolution in biomedical research, not only in the way research is conducted, but also in the way it is supported. While much progress has been made in the treatment and understanding of disease, it is becoming clear that in order to continue making important advances we will have to begin to investigate and decode the complexities associated with the disease process using holistic/systems level approaches while applying new insights by taking into account the specific genetic and environmental context of the individual patient.

Current reductionist approaches have provided technological and scientific advances and have helped set the stage for a new systematic approach in medical research. Starting with genomics, there has been an integration of high-throughput technologies (including microarrays, proteomics, new molecular and cellular imaging) into mainstream biology. The influx of large amounts of data and the associated management and analysis needs have naturally created a fertile ground for the application of methods from computational and mathematical sciences. As a result, we see today in many institutions diverse disciplines being increasingly integrated into all aspects of biomedical research. The promise is that many of the approaches, technologies, and thinking, previously separate from the traditional biomedical community, will now lend their strengths to many of these complex problems.

This change in investigational methodology has paralleled changes in the administration and funding of biomedical science. Beginning in May 2002, the National Institutes of Health (NIH), under the leadership of Elias A. Zerhouni, M.D., convened a series of meetings to chart a "roadmap" for medical research in the 21st century. The purpose was to identify major opportunities and gaps in biomedical research that no single institute at NIH could tackle alone but that the agency as a whole must address to make the biggest impact on the progress of medical research. Many of these efforts targeted some of the key challenges faced in deciphering disease complexity, along with opportunities for new discoveries. NIH is uniquely positioned to catalyze changes that must be made to transform our new scientific knowledge into tangible benefits for people. Developed with input from meetings with more than 300 nationally recognized leaders in academia, industry, government, and the public, the NIH Roadmap (http://nihroadmap.nih.gov/) provides a framework of the priorities that NIH, as a whole, must address in order to optimize its entire research portfolio. It lays out a vision for a more efficient and productive system of medical research.

The initial NIH Roadmap identified the most compelling opportunities in three main areas: (1) new pathways to discovery, (2) research teams of the future, and (3) re-engineering the clinical research enterprise. One of the most ambitious visions to come out of the roadmap process was a program to expand the computational infrastructure and software tools needed to advance biomedical, behavioral and clinical research. At the core of this effort are seven National Centers for Biomedical Computing (NCBC, http://www.bisti.nih.gov/ncbc/). The National Centers for the Multi-Scale Analysis of Genetic and Cellular Networks (MAGNet) is one of those centers. The centers, each funded at nearly $20 million over five years, are

part of a coordinated effort to build the computational framework and resources that researchers need to gather and analyze the massive amounts of biomedical data currently being generated by labs and clinics. This infrastructure will help the research community translate their data into knowledge that ultimately improves human health. Centers are dynamic partnerships of various research disciplines including computer scientists, biologists, engineers, and clinicians. To maintain the focus of the centers on current problems in biomedical research, each center has identified biological projects to drive the computational efforts and solidify multi-disciplinary teams. To further expand the impact of these centers the NIH has established roadmap related programs for Collaborations with the National Centers for Biomedical Computing (PAR-07-249 and PAR-07-250). These announcements invite applications from investigators to work on projects that broaden a center's biological or computational strengths. In their brief history, the NCBCs have established themselves as leading centers of research in bio-computing as well as a national resource for the greater research community.

"This infrastructure will help the research community translate their data into knowledge that ultimately improves human health."

While the NCBC initiative, as a roadmap activity, is a trans-NIH program, many individual institutes have also recognized and invested in the area of systems and computational biology. The National Cancer Institute (NCI), through the Integrative Cancer Biology Program (ICBP, http://icbp.nci.nih.gov/) has recently established a number of national centers to focus these efforts in the area of cancer biology. The NCI currently funds 9 ICBP centers focused on various aspects of cancer biology. Like the NCBCs, the center teams are composed of researchers with diverse scientific backgrounds. The goal of the ICBP is to use computational and experimental techniques to develop and apply predictive computational models describing various transforming processes of cancer. These models will prove essential in our eventual understanding and management of this disease, as well as in applications of personalized treatment. The ICBP has already established promising approaches for predicting signaling pathways, as well for 3-D tumor modeling. Critical to the success of both the ICBP and the NCBC programs is the establishment of a strong educational and outreach effort. This is not only important for the dissemination of the information and models; it is also critical for the training and education of young researchers in this emerging field.

MAGNet and the other NCBCs along with specific programs such as the ICBP, bring the needed resources and approaches to help understand and manage some of our most complex and deadliest diseases. These efforts will help enable the NIH and the biomedical community to sustain its historic record of making cutting-edge contributions that are central to extending the quality of healthy life for people in this country and around the world.

1. E. Zerhouni, Medicine. The NIH Roadmap. 2003, Science.;302(5642):63-72

2. Morris RW, Bean CA, Farber GK, Gallahan D, Jakobsson E, Liu Y, Lyster PM, Peng GC, Roberts FS, Twery M, Whitmarsh J, Skinner K. Digital biology: an emerging and promising discipline. Biotechnol. 2005 Mar;23(3):113-7.

3. Stilwell JL, Guan Y, Neve RM, Gray JW. 2007. Systems biology in cancer research: genomics to cellomics. Methods Mol Biol.;356:353-65.

4. Hornberg JJ, Bruggeman FJ, Westerhoff HV, Lankelma J., 2006. Cancer: a Systems Biology disease. Biosystems. Feb-Mar;83 (2-3):81-90.

MAPPING THE TRANSCRIPTION FACTOR MODULATOR REPERTOIRE IN HUMAN B LYMPHOCYTES

ANDREA CALIFANO, PHD RICCARDO DALLA-FAVERA, MD

Technical advances that enable monitoring the concentration of vast numbers of messenger RNAs, using microarray expression profiles, have greatly improved our ability to dissect the cell's regulatory networks. While these approaches have been used mostly to dissect transcriptional networks in prokaryotes, such as E. coli (Gardner et al. 2003; Faith et al. 2007), or in lower eukaryotes, such as yeast (Segal et al. 2003), MAGNet investigators have recently introduced new information theoretic methods (ARACNE) to study these networks in human cells (Basso et al. 2005; Margolin et al. 2006; Margolin et al. 2006). Transcriptional interactions predicted by ARACNE have been biochemically validated in vivo, using Chromatin Immunoprecipitation assays, first for MYC and BCL6 targets in Human B cells and more recently for several other transcription factors (TFs) in a variety of additional cellular contexts. These include the validation of: MYC and Notch1 targets in T cells (Palomero et al. 2006), CREB targets in peripheral leucocytes, STAT3, BHLHB2, RUNX1, CEBPB, and FOSL2 targets in glioblastoma cells, and PBX19 targets in rat brain tissue (manuscripts in preparation). In all these cases, biochemical validation was successful in 70% to 90% of the tests, showing that computational inference methods are approaching the accuracy of experimental assays.

Unfortunately, while providing a wealth of novel information on transcription factor candidate targets and a low false positive ratio, algorithms like ARACNE only scratch the surface of the complexity of transcriptional regulation processes. There are two fundamental reasons for this. First, transcription factors do not operate in isolation but rather in concert with many other proteins (such as co-factors and chromatin modification enzymes) that mediate the efficiency of their binding, the recruitment of transcriptional and repression complexes, and the accessibility of the chromatin molecule (among others). Second, the activity of transcription factors is itself regulated by signal transduction events leading to the activation or degradation of transcription factors and their complexes. This is achieved through a variety of well-characterized post-translational modification events such as phosphorylation, acetylation, sumoylation, ubiquitination, etc.

One way to think "visually" about such processes is to assemble a graph where nodes represent genes or their byproducts and arrows between nodes represent their physical interactions. In such a representation, the direct regulation of a target gene (e.g., TERT) by a transcription factor (e.g., MYC), as

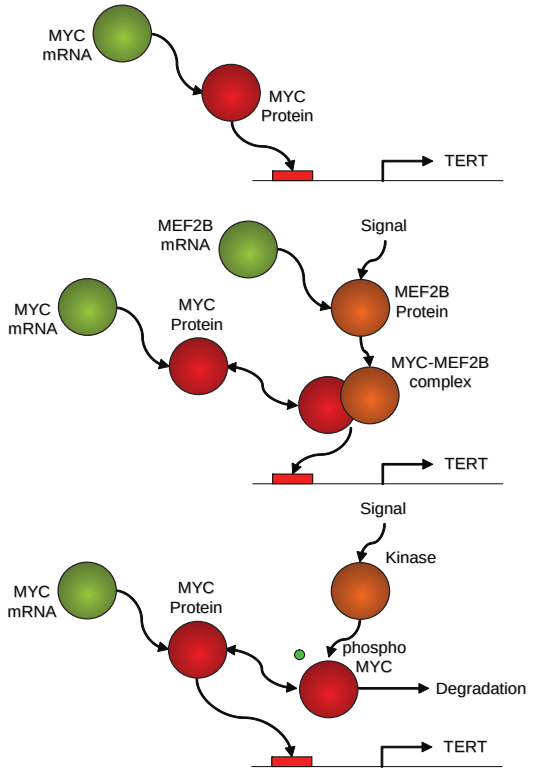


Fig. 1: Examples of graphical interaction network representations showing direct and modulated transcriptional interactions.

inferred by ARACNE, would be represented as a linear chain of individual physical interactions leading from the MYC mRNA to the TERT mRNA, see Fig. 1. This diagram provides a generic indication that the more MYC is expressed in the cell, the more TERT is likely to be expressed as well. When more complex interactions are considered, however, such as those involving post-translational modifications or complex formation, one must introduce additional nodes representing new transient or stable molecular species such as the phosphorylated version of a TF or a TF complex formed by two or more proteins. There are many proteins that affect the ability of MYC to regulate TERT, for instance, either specifically or non-specifically. An active kinase (such as GSK3), which destabilizes MYC by phosphorylation at Thr-58, will induce rapid degradation of the protein, thus reducing the gene's ability to regulate the expression of its targets, including TERT (Gregory et al. 2000). Conversely, as shown by the reporter gene assay in Fig. 2, the availability of the MEF2B co-factor will significantly increase the ability of MYC to activate TERT and a few other targets specifically, while leaving the majority of other MYC targets unaffected (Wang et al. 2007). The bottom half of Fig. 1 represents these more complex three-way interactions by introducing a new molecular species (i.e., either the phosphorylated version of MYC or the MYC-MEF2B complex).

Interestingly, while many algorithms are currently available to infer transcriptional targets of a TF, no algorithm has been proposed to systematically identify all the proteins that affect the ability of a TF to regulate some or all of its targets using gene expression profile (GEP) data. To some extent, the key obstacle to the development of such methods is relatively easy to understand. Since GEP data provides a snapshot, albeit a comprehensive one, of the mRNA in the cell, it is difficult to believe that it may provide evidence about interactions that occur exclusively at the post-translational (i.e., protein) level, such as the formation of TF complexes or a TF activation by phosphorylation.

This is precisely where the interdisciplinary background of MAGNet investigators is helpful. While most of the regulation of signaling proteins happens via signals, rather than transcriptionally, cell samples show some variability of the proteins at the mRNA level. For a biologist, the natural fluctuations of a specific gene's mRNA across individual cells and populations are a form of experimental noise, which hides the underlying biological information. If this sample-related variability could be reduced – the biologist would argue – cellular processes would be so much easier to dissect. However – a physicist would rebut, – as long as the natural sample variability is larger than the measurement error one should not be considered it as noise but rather as a physiologic (or pathologic) process that can be used to measure specific systems properties. Specifically, biological "noise" can be turned into signal if a sufficient number of samples are collected. In that case, if the cell is close to equilibrium or involved in dynamics that are slow compared to signaling processes, fluctuations in the expression of a gene across many samples can be used as a proxy for fluctuation in the corresponding protein availability. Furthermore, assuming that cellular signals are statistically independent of their substrate availability, a reasonable starting hypothesis, mRNA concentration data can then be used effectively to investigate post translational processes.

MINDY (Modulator Analysis by Network Dynamics) uses these principles to detect proteins that affect the transcriptional program of a TF, i.e., TF modulators. This can be best understood

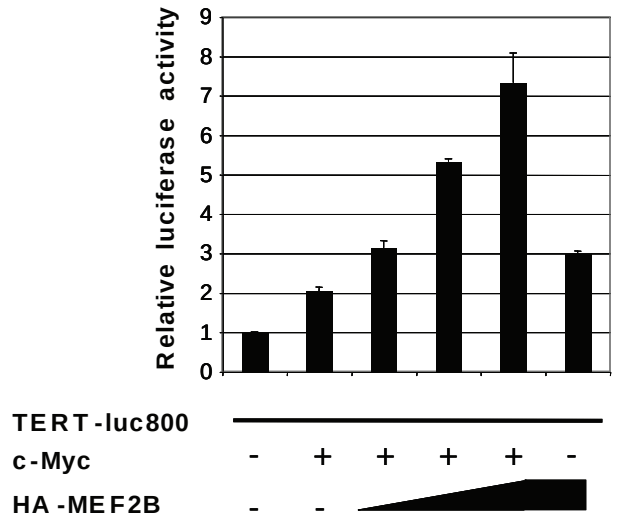


Fig. 2: TERT-luciferase assay results from T239 transfected T cells.

with an example: suppose that a kinase, K_s, activated by a signal S, is required to activate a TF in turn and to allow it to regulate its target(s) t. Then, a transcriptional fluctuation of the kinase mRNA, K_m, should result in a corresponding fluctuation in the amount of K_s and thus in the amount of active TF. Under this assumption, the Califano lab has shown that the conditional mutual information $I[TF; t \mid K_m]$ becomes a non-constant function of K_m. This can be efficiently assessed by showing that $\Delta I = |I [TF; t \mid K_m+] - I [TF; t \mid K_m-]|$, the absolute difference of the mutual information computed from the samples where K_m is most expressed (K_m+) and from those where it is least expressed

(K_m-), is greater than zero. This test is a necessary and sufficient one if the dependency of the conditional information on K_m is monotonic, which can be easily shown to be the case for realistic biochemical interactions.

Suppose for instance, as an extreme case, that K_m is completely absent from the cell (K_m-). Then, even if the signal S were present, the TF could never become active and thus $I [TF ; t | K_m] = 0$, because the TF cannot regulate any of its targets, including t. On the other hand, if the kinase were present in abundance (K_m+), any amount of signal S would find a sufficient K substrate to activate the kinase, which would in turn activate the TF and lead to target regulation. Under this condition the marginal information $I [TF ; t | K_m+] > 0$, because changes in TF would correlate with changes in target expression. Thus, trivially, $\Delta I > 0$. As the kinase availability range becomes narrower, as is the case in natural sample variability, the corresponding ΔI would also decrease. However, if we assume that the signal S is independent of the kinase availability (a reasonable starting hypothesis), then no matter how narrow the natural variability in kinase concentration range is, there will always exist a data sample size at which the corresponding change in information becomes statistically significant. Interestingly, as discussed later, even relatively modest sample sizes (N > 200), such as are available in today's GEP repositories can provide some valuable insight into the cell's post-translational interactions.

Hence, MINDY requires computing the difference in mutual information between the TF and a target t in two subpopulations, one where the candidate modulator gene m (our kinase, for

Modulator	M#	M+	M-	Mode	Description	Evidence
CSNK2A1	205	205	0	+	Casein kinase 2, alpha 1	HPRD
PPAP2B	120	0	120	-	Phosphatidic acid phosphatase 2B	Acitvates GSK3
HCK	118	0	118	-	Hemopoietic cell kinase	BCR Pathway
SAT	109	0	109	-	Spermidine N1-acetyltransferase	
DUSP2	95	0	95	-	Dual specificityphophatase 2	Desphosphorylates ERK2
MAP4K4	94	0	94	-	MAP kinase kinase kinase kinase 4	BCR Pathway
PPM1A	92	0	92	-	Proteinphosphatase 1A	
CSNK1D	90	0	90	-	Casein kinase 1, delta	
GCAT	86	86	0	+	Glycine C-acetyltransferase	
TRIO	84	0	84	-	Triple functional domain	
PRKCI	63	63	0	+	Proteinkinase C, iota	BCR Pathway
PRKACB	57	0	57	-	Proteinkinase, catalytic, beta	BCR Pathway
STE38	56	56	0	+	Serine/theronine kinase 38	
MTMR6	55	2	53	-	Myotubularin related protein 6	
NEK9	53	53	0	+	NIMA-related kinase 9	
MYST1	47	47	0	+	MYST histone acetyltransferase 1	
MAPK13	45	45	0	+	MAP kinase 13	BCR Pathway
OXSR1	45	0	45	-	Oxidative-stressresponsive 1	
DUSP4	43	0	43	-	Dual specificityphophatase 1	
MAP2K3	42	0	42	-	MAP kinase kinase 3	BCR Pathway
PPP4R1	39	0	39	-	Proteinphosphatase 4, R1	
ERK2	37	37	0	+	MAP kinase 1	BCR Pathway
MAP4K1	36	0	36	-	MAP kinase kinase kinase kinase 1	BCR Pathway
CSNK1E	35	34	1	+	Casein kinase 1, epsilon	
FYN	33	0	33	-	FYN oncogene	
NEK7	33	33	0	+	NIMA-related kinase 7	
CSNK2A2	31	31	0	+	Casein kinase 2, alpha	Related to CSNK2A1
DUSP5	30	0	30	-	Dual specifictiyphosphatase 5	

Table 1: Results of the MINDY analysis for MYC. Column 1 shows the modulator gene symbol; column 2 shows the number of affected MYC interactions; columns 3 and 4 show respectively the number of interactions that become more correlated with MYC when there is respectively an increased or decreased amount of the modulator gene; column 5 is inferred from 3 and 4 and indicates the modulation mode (+ = MYC activator, - = MYC antagonist); column 6 is a gene description and column 7 shows literature clues about MYC modulation. Blue genes are previously known in the literature to affect MYC function. Green genes were experimentally validated in the Califano/Dalla Favera lab.

instance) is least expressed and one where it is most expressed, from a relatively large sample set. If the absolute difference is deemed statistically significant given the sample size, after Bonferroni correction for the number of tests performed, then the gene m is considered a putative modulator of the interaction between the TF and the target t. Given a putative modulator, the test can be performed on each candidate target of the TF. These can either be selected among all the genes on the microarray expression profile or from a set of known TF targets (e.g., from the literature or from ARACNE). Table 1 shows the result of this procedure where MYC is the selected transcription factor, candidate modulators are chosen among all kinases, phosphatases, and acetyltrasferases on the GEP, and the targets of MYC are selected from all ChIP and ChIP-Chip validated MYC targets in the MYC target database (Zeller et al. 2003). Only the top modulators, affecting 30 or more MYC-target interactions are shown.

Remarkably, as shown by columns 3 and 4, even though each set is performed in isolation, there is complete consistency across the different tests for each putative modulator. For instance, Casein Kinase 2 (a protein known to phosphorylate MYC and to stabilize the MYC-MAX heterodimer) is found to be the most statistically significant MYC modulator, affecting 205 of the ~340 tested MYC target interaction. As shown, all such interactions demonstrated an increase of mutual information when there was more Casein Kinase 2 (see counts in column 3), consistently with the known role of this protein. None of the MYC target interactions became more correlated (increase in mutual information) when there was less Casein Kinase 2 in the cell (see 0 count in column 4). The opposite is true for proteins that act as MYC antagonist (e.g. PPAP2B) where all the counts are in column 4 rather than 3. This behavior is reflected across all the reported putative modulators, showing that the analysis produces biologically plausible results.

Similar results are also obtained for candidate modulators that are transcription factors, leading to the dissection of combinatorial regulation programs.

Further extensive biochemical validation of several modulators, using co-IP, reporter gene assays, and modulator silencing by siRNA, shows that the method is capable of identifying novel post-translational modulators of the MYC protein, including signaling proteins, such as STK38 and HDAC1, and co-factors such as

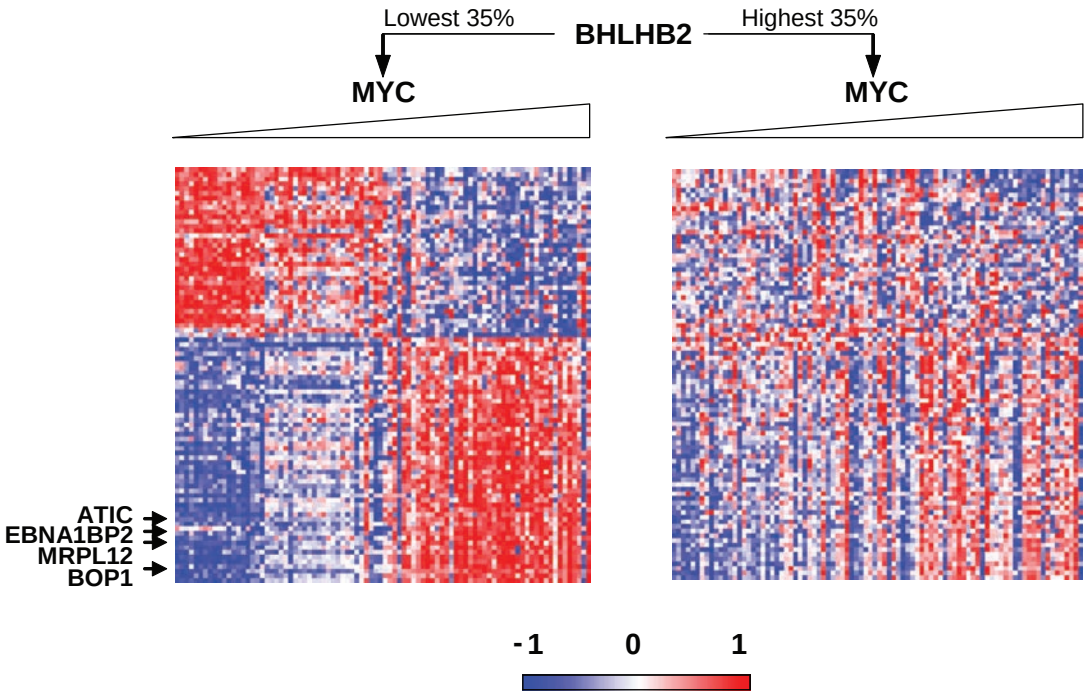


Fig. 3: BHLHB2 analysis. The figure shows how MYC appears to regulate its targets in the samples with the lowest concentration of BHLHB2 mRNA, while regulation of the same targets appears to be lost in the samples with a substantial amount of BHLHB2 mRNA.

BHLHB2 and MEF2B. For instance, Figure 3 shows differential regulatory ability by MYC in the presence or absence of BHLH2. This was confirmed by a TERT-luciferase reporter gene assay, showing a decrease in TERT expression as BHLHB2 levels are increased in the cell.

MINDY was run for each signaling and TF protein against each TF protein, to dissect both the interface between signal transduction and transcriptional regulation as well as the combinatorial nature of transcriptional programs. Validation, using existing pathways, shows that 30% to 70% of the inferred modulators are in the same pathway as the TF they affect, depending on the minimum number of affected targets. This is providing the first genome-wide and systematic analysis of all post-translational modulators of every TF in a B Cell. This information can be used both for the identification and validation of therapeutic targets, as well as for the dissection of pathways that are dysregulated in lymphoid malignancies.

1. Basso, K., A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera and A. Califano (2005). "Reverse engineering of regulatory networks in human B cells." Nat Genet 37(4): 382-90.

2. Faith, J. J., B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins and T. S. Gardner (2007). "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles." PLoS Biol 5(1): e8.

3. Gardner, T. S., D. di Bernardo, D. Lorenz and J. J. Collins (2003). "Inferring genetic networks and identifying compound mode of action via expression profiling." Science 301(5629): 102-5.

4. Gregory, M. A. and S. R. Hann (2000). "c-Myc proteolysis by the ubiquitin-proteasome pathway: stabilization of c-Myc in Burkitt's lymphoma cells." Mol Cell Biol 20(7): 2423-35.

5. Margolin, A. A., I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera and A. Califano (2006). "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context." BMC Bioinformatics 7 Suppl 1: S7.

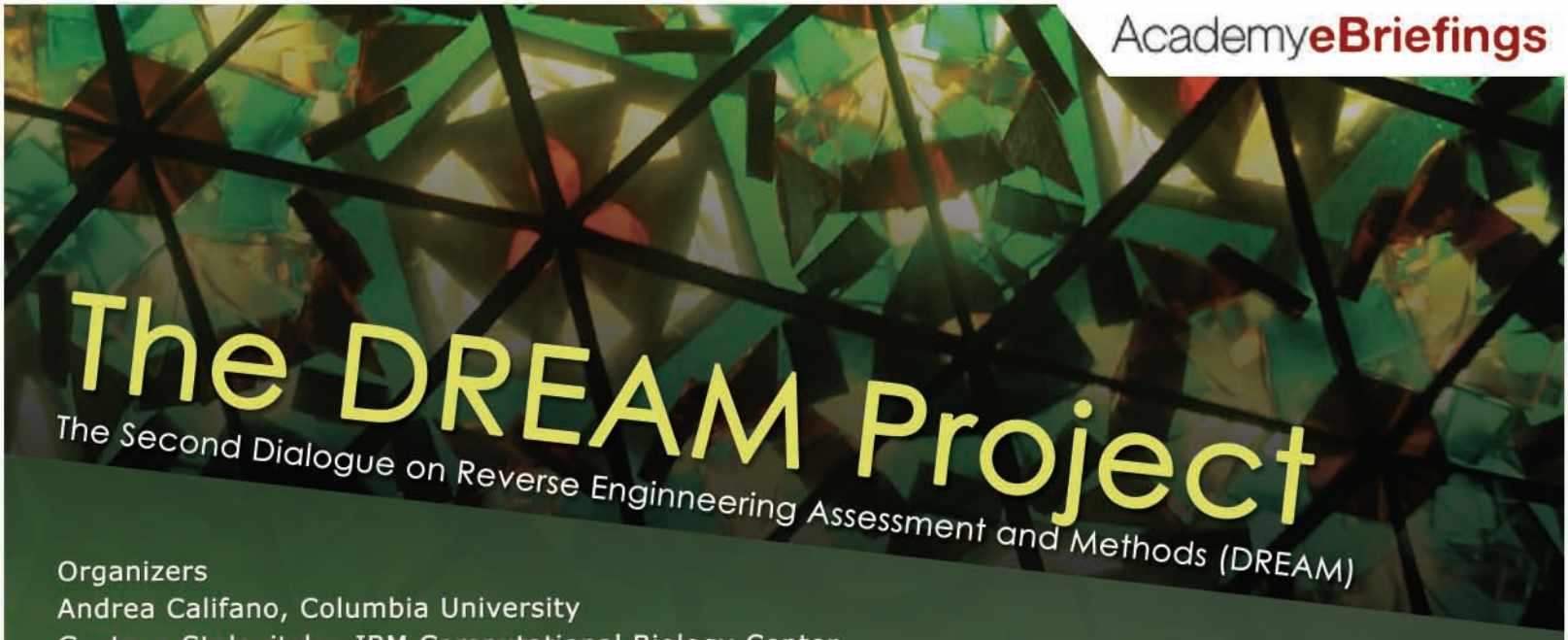
6. Margolin, A. A., K. Wang, W. K. Lim, M. Kustagi, I. Nemenman and A. Califano (2006). "Reverse engineering cellular networks." Nat Protoc 1(2): 662-71.

7. Palomero, T., W. K. Lim, D. T. Odom, M. L. Sulis, P. J. Real, A. Margolin, K. C. Barnes, J. O'Neil, D. Neuberg, A. P. Weng, J. C. Aster, F. Sigaux, J. Soulier, A. T. Look, R. A. Young, A. Califano and A. A. Ferrando (2006). "NOTCH1 directly regulates c-MYC and activates a feed-forward-loop transcriptional network promoting leukemic cell growth." Proc Natl Acad Sci U S A 103(48): 18261-6.

8. Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller and N. Friedman (2003). "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data." Nat Genet 34(2): 166-76.

9. Wang, K., M. Saito, I. Nemenman, K. Basso, A. A. Margolin, U. Klein, R. Dalla Favera and A. Califano (2007). "Genome-wide identification of transcriptional network modulators in human B cells." submitted to Nature.

10. Zeller, K. I., A. G. Jegga, B. J. Aronow, K. A. O'Donnell and C. V. Dang (2003). "An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets." Genome Biol 4(10): R69.



The DREAM Project

The Second Dialogue on Reverse Enginneering Assessment and Methods (DREAM)

Organizers
Andrea Califano, Columbia University
Gustavo Stolovitzky, IBM Computational Biology Center

Sponsored by the Systems Biology Discussion Group

The Systems Biology Discussion Group is made possible with generous support from Columbia University, IBM, Merck, Pfizer, NIH, Roadmap and MAGNet

Held at the New York Academy of Sciences | Dec 3-4 2007
http://www.nyas.org/dream2007

The DREAM

At a microscopic level, organisms are ruled by interacting systems of biomolecules. Historically, scientists painstakingly elucidated chains of molecular events using experiments that reveal individual interactions, although they recognized that members of different pathways frequently interact. In recent years, researchers have built richer, interconnected networks to mathematically summarize their knowledge of these interactions. This systems biology enterprise, largely stimulated by high-throughput tools like microarrays that measure mRNA levels as an indicator of gene expression, is a vital and increasingly important activity in both basic biology and in medicine.

A nagging concern, however, is how accurately these networks represent the biology. For complex systems like biological networks, there are practical limits on how well even massive amounts of data can uniquely define the underlying structure and yield useful predictions of measurable events. Indeed, although its advocates call this process “reverse engineering,” the topology and the detailed molecular interactions of the “inferred” networks will likely never be known with precision.

On December 3 and 4, 2007, the New York Academy of Sciences hosted the second meeting of the Dialogue on Reverse-Engineering Assessment and Methods (DREAM), which the Academy has nurtured from its inception. (For more information, see the related volume of the Annals of the New York Academy of Sciences: Reverse Engineering Biological Networks: Opportunities and Challenges in Computational Methods for Pathway Inference.) This ongoing process aims to assess the ability of scientists—and their computer servants—to infer networks from experimental data, by comparing their predictions to “gold-standard” networks whose structure is thought to be known. The conference also featured plenary and invited talks, as well as contributed talks and posters, illuminating various aspects of the reverse-engineering challenge.

Diverse networks

The centerpiece of the second DREAM meeting was a set of five “challenges,” in which participants tried to replicate various types of known networks from specified data. The five challenges included identifying targets of the transcriptional repressor BCL6, determining

continued on next page...

continued...

which proteins of a group interact, and inferring the topology of a variety of networks, including a five-gene synthetic network in yeast, several more complex, computer-generated networks, and a documented gene regulatory network in a bacterium.

To ensure a fair comparison of different techniques for reverse engineering networks, the DREAM organizers carefully limited the data supplied, and tried to disguise it so that participants could not leverage other kinds of data. This blinded procedure does not take advantage of all available information, however, especially biological wisdom that does not fit easily into a formal mathematical framework. Some speakers instead advocated incorporating prior biological knowledge such as known feedback loops into the network from the earliest stages of the process. But others felt that, although such information might improve the networks, it would compromise the primary DREAM goal of assessing methods.

Determining the most revealing experimental conditions is a crucial issue for reverse engineering. The blinded competition, however, demanded that the organizers provide the data, so the competitors could not differentiate themselves by devising perturbations to best clarify network features.

Transcriptional regulation—in which proteins produced from mRNA in turn act to modulate the transcription of other genes into mRNA—is the poster child of systems biology. Researchers exploit uniform and commercially accessible high-throughput data to construct complex transcriptional networks based on simple models of regulation. Nonetheless, recent studies reveal important complexities in transcription regulation. In addition, other types of interaction must ultimately be integrated into the description. Researchers have made significant progress in elucidating some types of networks, such as signaling networks driven by post-translational modifications of proteins. Other networks, like those governed by metabolic interactions or the various mechanisms associated with microRNA, are at an earlier stage of understanding.

Diverse algorithms

The purpose of DREAM is not to produce the best possible network, but to evaluate the best tools for producing networks. The choice of tools depends in part on the nature of the available data. Dynamic techniques aim to exploit the detailed time evolution of biological responses like mRNA concentration in response to perturbations. The underlying model is generally a system of differential equations, and the modeling aims to determine the parameters of these equations.

Many algorithms analyze the correlations between the steady-state levels of biomolecules, such as mRNA, under various conditions. These static techniques use statistical methods to try to distinguish the direct interactions between nodes from those mediated by other nodes. Their results are generally embodied in the topology of a (possibly directed) graph.

For both static and dynamic models, however, the experimental data are typically insufficient to specify a unique network. Researchers generally must discard many apparent interactions because their effects are unimportant, but in so doing they may also discard some real interactions. Developing metrics that quantify this tradeoff is a subtle and challenging issue, especially for biological networks, which are often sparse.

Diverse results

In the end, 36 teams made a total of 110 predictions for the five challenges. The match between these predictions and the “known” networks varied widely, both between teams and between challenges. For example, all teams did poorly at identifying the most complex in silico network, which was governed by transcriptional, signaling, and metabolic interactions. The networks inferred from the data differed significantly from the real network, which is precisely known. What is not known is whether the data given are, by themselves, sufficient to distinguish the networks.

By contrast, many teams did very well at identifying the targets of the transcription suppressor BCL6 from expression and sequence data. For this real data, however, the “gold-standard” result is itself derived in the context of a specific understanding of the biological mechanisms. Although the organizers did additional experiments to validate the results, the team that best predicted the targets used hints about the organizers’ thinking process to better tune their predictions. They challenged the organizers to consider that, rather than identifying the underlying network, predicting the observable results of experiments may be a more objective way to asses reverse engineering.

At this early stage, the DREAM process is still searching for the best ways to find networks, and each challenge has shed some light on the the problem.

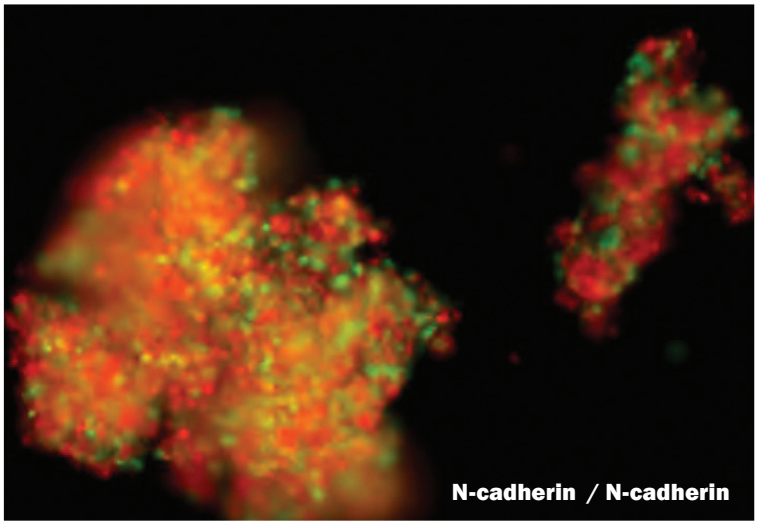
UNDERSTANDING CADHERIN SPECIFICITY IN THE DEVELOPMENT OF MULTICELLULAR STRUCTURES: A COMBINED EXPERIMENTAL AND COMPUTATIONAL STUDY

LAWERENCE SHARIPO, PHD

DEPARTMENT OF BIOCHEMISTRY AND BIOPHYSICS
COLUMBIA UNIVERSITY

BARRY HONIG, PHD

DEPARTMENT OF BIOCHEMISTRY AND BIOPHYSICS
CENTER FOR COMPUTATIONAL BIOLOGY AND BIOINFORMATICS
COLUMBIA UNIVERSITY



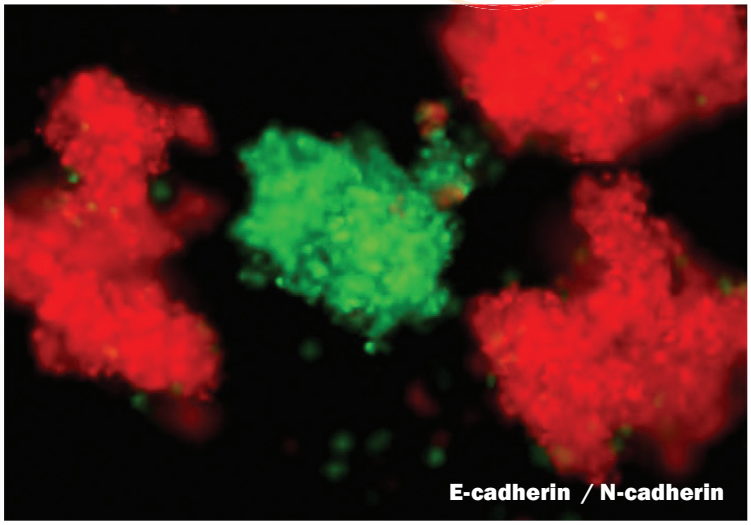


Fig. 2: Cell separation mediated by N- and E-cadherins recapitulated in transfected cells.

Cadherins (calcium dependent adherent proteins) comprise one of the largest families of cell surface adhesion proteins. Their regulated expression, in different cells at different times during development, guides the formation of specific multicellular structures. There are about 100 different cadherins in the human genome, in five different large subfamilies. The members of each subfamily are highly related to one another.

In the context of one of the MAGNet Center's driving biological problems, our laboratories have come together to study cadherins by combining physicochemical and computational investigations. Our goal is to understand the molecular basis of the binding specificity of cadherins and, in turn, the structural and energetic basis of many cell-cell adhesion processes. The fundamental question we ask is how cadherins are able to bind to one another with sufficient specificity to accomplish their cell recognition function, even though

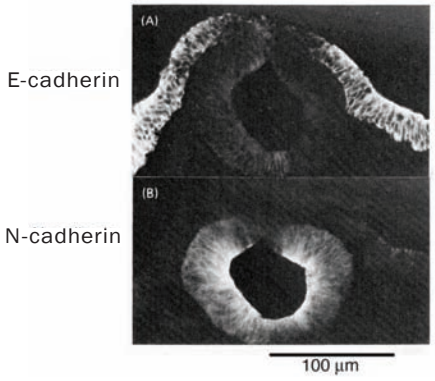


Fig. 1: The process of neurulation, common to all vertebrates, is driven by regulated changes in expression of E- and N- cadherins. This micrograph, from Masatoshi Takeichi's laboratory, shows a slice through a 6-day post fertilization chick embryo.

many cadherins are closely related to one another in sequence, and thus might be expected to cross-react. In addition, we wish to understand the diverse function of different cadherins and to see how different cadherin subfamilies have evolved to carry out distinct, albeit related functions.

A classic example of cadherin function can be seen in embryonic tissue development where cells in the neural tube that express N-cadherin separate from epithelial cells that express E-cadherin (Figure 1). This phenomenon can be replicated in in-vitro cell assays which show that cells transfected with N and E cadherin sort out from one another

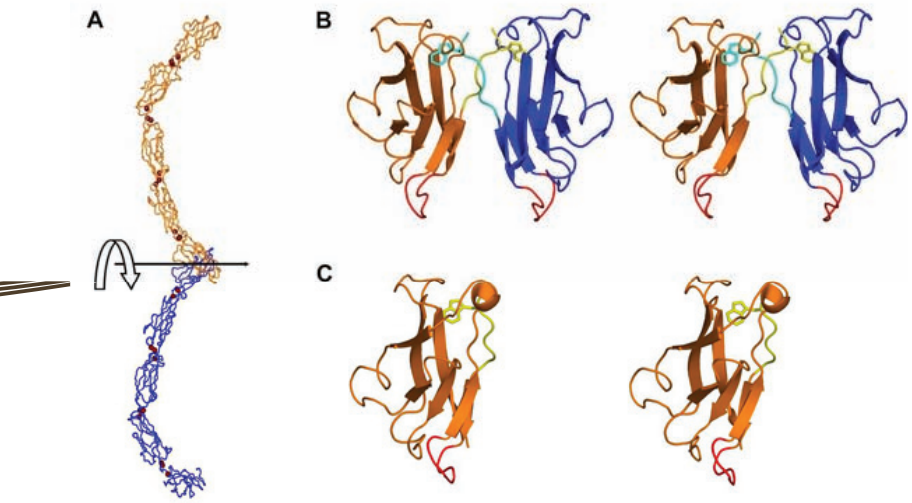


Fig. 3: Structural models of C-cadherin. a) The crystal structure of the entire ectodomain of C-cadherin, determined in our lab. b) Structure of the EC1 domain dimer from C-cadherin (stereo diagram). The swapped A strands, including the conserved Trp-2 side-chain, are shown in yellow and cyan. The putative hinge loop is shown in red. c) A homology model of the monomeric form of the C-cadherin EC1 domain based on the structure of the E-cadherin monomer (PDB: 1O6S). The A-strand is shown in yellow with the Trp-2 side-chain facing the interior of its own protomer. The hinge loop is shown in red. Dynamic exchange between monomer and dimer is a critical feature of cadherin adhesion.

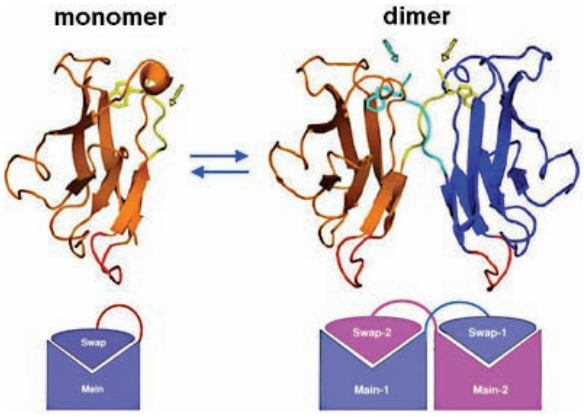


Fig. 4: Monomer (PDB 1FF5) and dimer (PDB 1L3W) structures of classical cadherin EC1 domains, and schematic diagram of domain swapping mechanism. 3D domain swapping, in general, requires a protomer consisting of a "main" domain and a "swapped" domain connected by a flexible hinge loop. In this way, symmetric dimmers can be formed simply by changing the conformation of the hinge loop. All molecular contacts between "main" and "swapped" domains are locally identical in the dimer and monomer forms, except that they are intramolecular in the monomer, and intermolecular in the dimer. The A-strand, containing Trp2 and shown in yellow, constitutes the "swapped" domain for classical cadherins. Since the A strand can bind to the body of its own protomer, classical cadherins effectively carry their own competitive inhibitors, and this is critical to their binding specificity.

into separate aggregates (Figure 2).

Cadherin adhesive dimers form through a strand-swapping mechanism (Figure 3), a specific type of the more general "domain swapping" phenomenon (Figure 4). We have shown, through a theoretical analysis, that this mechanism imparts novel energetic properties to cadherins, enabling high specificity while maintaining low affinity (Figure 4). Low-affinity binding is a requirement for cadherins because they function as membrane attached "lawns" of proteins that bind cell surfaces together. High affinity interactions would hold cells together permanently, and impede the dynamics of development. Low-affinity binding is a characteristic of most or all cell adhesion proteins. The domain swapping mechanism may prove to be a general mechanism used by other families of cell adhesion proteins as well.

In order to address the question of how cadherins sort cells into tissue layers, we are taking a two-pronged approach. First, we are using surface plasmon resonance – a tool to experimentally measure binding strength and kinetics – to characterize the binding energies of cadherin pairs (Figures 5 and 6). Second, we are developing theoretical models of cell sorting, based on the idea, originally proposed by Malcolm Steinberg at Princeton, that cell aggregates behave as viscous liquids, and the equilibrium configuration of cell assemblies will depend on interaction energies between cells. These cell-level interaction energies are determined by the interaction strength and number of adhesion molecules on the cell surface.

Results from SPR experiments (Fig. 6) show that cadherins bind in the micromolar range: N-cadherin homodimerizes with KD ~ 20μM and E-cadherin homodomerizes with much weaker affinity, about 80 μM. Very surprisingly, we have found that the binding strength of the heterophilic E-cadherin/N-cadherin interaction is intermediate between these two values.

These data suggest the need to reinterpret

our understanding of neurulation – the E-cadherin and N-cadherin mediated separation of the neural tube from the ectoderm (Figure 1). Although E-cadherin expressing cells bind together and N-cadherin cells bind together, it was unexpected to find that the adhesion molecules can also interact heterophilically.

A simple theoretical analysis of cell sorting, however, clears up this apparent paradox (Fig. 7), and shows that these binding affinity results can beautifully explain the observed cell layer separation.

Just as oil separates from water based on the relative homophilic (water-water and oil-oil) and heterophilic (oil-water) binding energies, cells apparently separate into tissues according to similar rules. For the case in which heterophilic binding is intermediate between the two homophilic binding energies, for a set of bounded "phases", it is predicted that the high-affinity cells (corresponding to N-cadherin) will form a core enveloped by a phase made from the low-affinity cells (corresponding to E-cadherin expressors).

These results provide a beginning glimpse into the effects of cell adhesion on the mechanics of tissue development. Many tasks remain including determination of the binding energies for all 19 classical cadherins conserved in vertebrate genomes, and mapping these to expression patterns at critical developmental stages in animals.

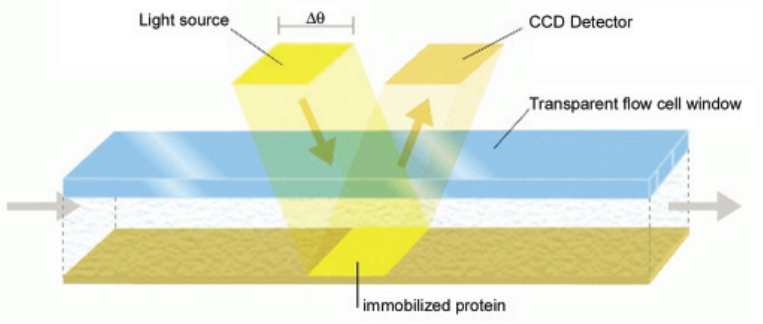


Fig. 5: Experimental design of SPR experiments. A protein is immobilized on the surface of a gold chip. A protein solution is flowed over the chip, enabling binding. The SPR angle, t, depends only on the amount of mass bound to the chip surface. Thus, binding interactions are detected as changes in this angle.

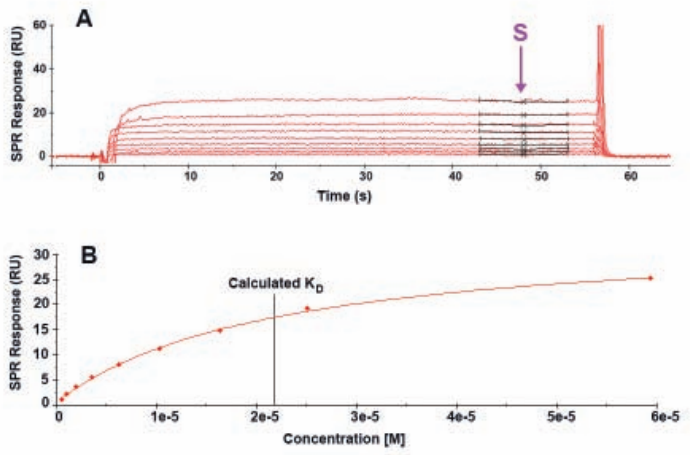


Fig. 6: (A) SPR traces of a concentration series (0.565-60.0 μM) of N-cadherin analyte, injected over a neutravidin Biacore chip coated with biotinylated N-cadherin. SPR at a steady-state point (S) are plotted in (B) and fit to a 1:1 binding model, yielding KD= 21.8± 1.4 μM.

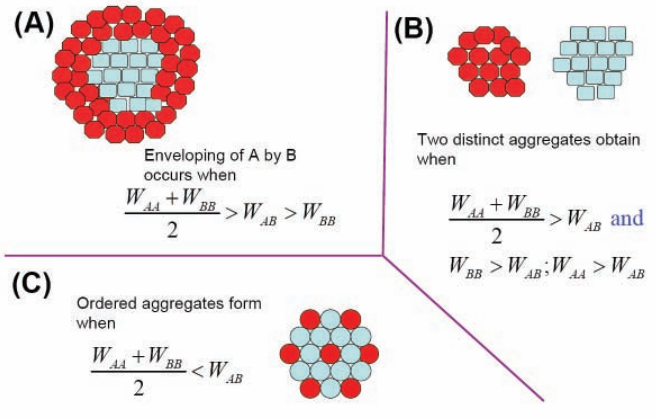


Fig. 7: Theoretical modeling predicts the outcome of sorting experiments for cells expressing equal levels of either of two cadherins – A (light blue) and B (red). Three different outcomes are predicted, which depend on the Work (W_{ij}) required to "pull

XPLORIGIN: A SOFTWARE FOR DECIPHERING POPULATION OF ORIGIN DEVELOPED AT THE PE'ER LAB

ITSIK PE'ER LAB

Despite our obsessive interest in humans, they make a poor model organism. Their genetics, for example, is complicated by generations of sorting into populations and merging them together. These violations of standard, statistical assumptions of random mating, idealized samples are a major problem in disease association studies. Fortunately, the information in genome wide arrays that profile an individual's genetic makeup for disease studies also stores clues about origin of an individual's ancestors. Like white light being separated into its constituent spectrum of colors, an individual's genetic variation can be better understood when decomposed into the ancestry backgrounds of that individual.

The Pe'er lab has recently completed development of Xplorigin (http://www.cs.columbia.edu/~itsik/Xplorigin/Xplorigin.htm) a software tool to decipher population ancestry of different regions along an individual's genome. This tool was used to analyze admixture in the population of Kosrae, Micronesia, in a genome wide association study of the Metabolic Syndrome. Xplorigin is based on a Generalized Hidden Markov Model, trained on data from the International HapMap Project (http://www.hapmap.org/). Further development of this tool is currently under way to allow statistical interpretation of genetic association studies in admixed population, taking into account this decomposition into ancestral origin population.

MUTAGENESYS - DIAGNOSTIC PREDICTIONS BASED ON GENOTYPE DATA

KENNETH ROSS AND ITSIK PE'ER LABS

MutaGeneSys is a new system developed by Julia Stoyanovich in the Ross lab, in joint work with Itsik Pe'er. This system uses genome-wide genotype data for disease prediction. MutaGeneSys integrates three data sources: the International HapMap project (http://www.hapmap.org), whole-genome marker correlation data and the Online Mendelian Inheritance in Man (OMIM, http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim) database. It accepts SNP data of individuals as query input and delivers disease susceptibility hypotheses even if the original set of typed SNPs is incomplete. The system is scalable and flexible: it operates in real time and produces population, technology, and confidence-specific predictions.

MutaGeneSys allows detection of individuals susceptible to OMIM disorders among participants of whole genome association studies, a yet unexplored perspective of such data. This system and its successors will pave the way for using whole genome SNP arrays as practical diagnostic tools. The findings of MutaGeneSys are currently being incorporated into the HapMap Web Browser as the OMIM_Associations track.

You can learn more about the MutaGeneSys Project at:

http://www.cs.columbia.edu/~jds1/MutaGeneSys/

IDENTIFYING GENE-PHENOTYPE ASSOCIATIONS IN HUMAN B LYMPHOCYTES

ANDREA CALIFANO LAB

The accurate reconstruction of networks of cellular interactions has provided valuable insight into the mechanisms that underlie normal and pathogenic processes. As our knowledge of these networks evolves, they can begin to be used as tools to further characterize disease on a genome-wide scale. In particular, we can identify how specific changes in the network are related to specific cellular phenotypes, and whether these changes can be traced back to a specific causal event.

With this context in mind, we have developed a systems biology approach to identify gene-phenotype associations in human B lymphocytes. Using a Bayesian evidence integration scheme, we have generated a comprehensive network of interactions present in B cells, as evidenced from various sources including literature mining, reverse engineering algorithm (ARACNE, MINDY), expression profiling, and databases such as BIND, IntAct and TransFac. A unique characteristic of this network is that it is hybrid in nature, including protein-protein interactions (PPI), protein-DNA or regulatory interactions (PDI), and higher-order modulated interactions (MI) in which a transcription factor and a target have a relationship dependent upon the expression level of a third modulator gene. The inclusion of all of these interaction types allows this B Cell Interactome (BCI) to cover a far greater extent of real relationships present within a typical B cell.

The analysis uses the concept that a particular gene, which is causally related to a specific phenotype, will show a pattern of changes in the network which can be identified by looking at its behavior with respect to its interaction partners. Using a large compendium of B Cell expression profiles covering over 20 normal and malignant phenotypes, we find edges in the BCI that show a gain-of-correlation (GOC) or loss-of-correlation (LOC)

pattern in a particular phenotype of interest. In other words, these are interactions which appear to be correlated in one phenotype but not in any other, and vice-versa; they are identified by looking at the change in correlation when a particular phenotype is removed. By grouping these modified interactions together, we can see which genes show a high enrichment in these changes, indicating they are behaving differently in that phenotype. We can score them as more likely to be a key causal gene (e.g. an oncogene in a tumor phenotype) or a key effector of the phenotype transition.

Results have shown promise in identifying key causal mechanisms. In 4 phenotypes (1 normal, 3 cancer), this method identified the known causal gene in the top 0.3% of all candidate genes. Moreover, the top lists for these phenotypes included several genes known to be active in or related to these phenotypes. For example, the MYC proto-oncogene was identified as a key gene involved in Burkitt Lymphoma (BL), where it is known to be translocated from chromosome 8 and aberrantly expressed. Also present however was MTA1, which has been shown to be necessary for MYC to hold its transforming capability. What makes these findings more interesting is that they would not have been identified by simple differential expression analysis in 3 out of 4 cases, indicating that a comprehensive systems-based approach can yield more insight than conventional approaches.

This method is currently being applied to further characterize more heterogeneous phenotypes, such as Diffuse Large B-Cell Lymphoma (DLBCL) and Chronic Lymphocytic Leukemia (B-CLL). We are hopeful that our comprehensive, evidence integration approach can be used to identify novel candidate genes involved in development of these lymphomas.

PROTEIN DATABASE CREATED USING NEW PIPELINE METHOD

BURKHARD ROST LAB

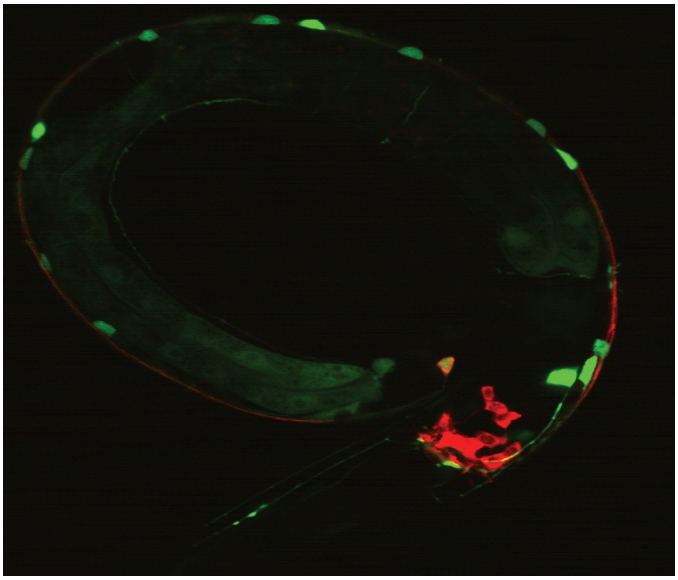
For transmembrane proteins, the presence of the lipid bilayer produces an amphipathic environment for individual strands or helices, a feature often detectable in amino acid sequences. To quantify this effect, we developed a pipeline that automatically identifies all lipid-facing, buried, and water-facing atoms in TM proteins, using the lipid bilayer position estimated from the Orientations of Proteins in Membranes (OPM, http://opm.phar.umich.edu/) database. Further, using our refined definition for local helix or strand axis, we annotate each residue with the periodic angular position relative to the face of maximal lipid exposure for each TM helix or strand. The resulting database consists of 92 TM alpha helical proteins containing 1465 transmembrane helices (295 sequence-unique), while the compilation of TM beta barrels is still underway. Such a database is likely to be useful as training data for diverse sequence-based prediction approaches.

IDENTIFYING THE BIOMOLECULAR PATHWAYS UNDERLYING SYNAPTIC CONNECTIVITY IN NEMATODE C. ELEGANS

DIMITRIS ANASTASSIOU LAB

The nematode C. elegans has a well-defined nervous system with only 302 neurons interconnected according to a known wiring diagram. If we also know the expression profiles of the individual neurons, we are presented with a unique opportunity to link the "single-neuron transcriptome" with the wiring diagram, identifying genes that are jointly associated with the presence of synapses, thus providing valuable help for the solution of the important effort of identifying the biomolecular pathways underlying synaptic connectivity.

This kind of research is at the heart of MAGNet's central theme (multiscale genomic and cellular networks), because it achieves the integration of interactions through two levels of abstraction, (a) the intercellular level of the neural interconnection network



Highlighting Neurons in Nematode C. Elegans: Using technology based on fluorescent proteins, we can see and isolate individual neurons in C. elegans - courtesy David Miller

and (b) the intracellular level of the biomolecular network within each of the neurons.

The single-neuron transcriptome of C. elegans is not yet known. We are collaborating with the laboratory of Prof. David Miller at Vanderbilt University, who uses pioneering cell-sorting and microarray-based technologies to profile mRNA isolated from individual neurons, gradually expanding our knowledge. Using the limited existing knowledge, we already have some preliminary results [1], and we are currently using novel computational techniques that we developed [2] to identify sets of genes that are synergistically interacting with respect to synapse formation.

[1] V. Varadan, D. Miller III and D. Anastassiou, "Computational Inference of the Molecular Logic for Synaptic Connectivity in C. elegans," Bioinformatics, Vol. 22, Issue 14 – ISMB 2006, pp. e497-e506, July 2006.

[2] D. Anastassiou, "Computational Analysis of the Synergy among Multiple Interacting Genes" (Review Article), Molecular Systems Biology, Vol. 3, No. 83, February 2007.

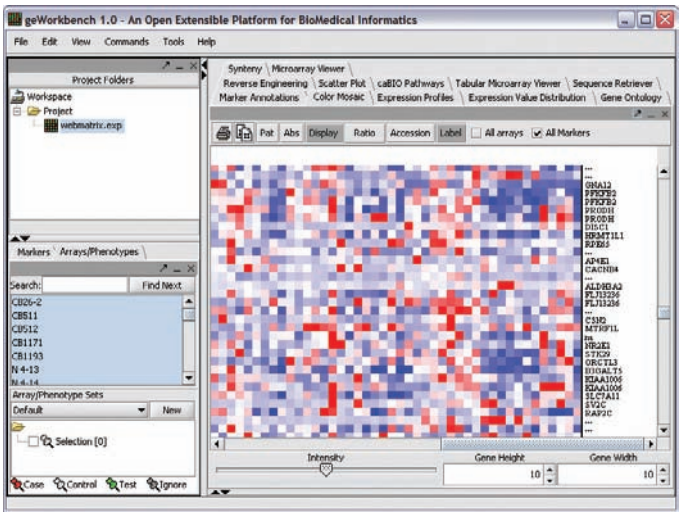
MAGNET CENTER TOOLS – PULLING EVERYTHING TOGETHER

ARIS FLORATOS AND ANDREA CALFANO LABS

An important mandate for all the National Centers for Biomedical Computing is to develop technologies and mechanisms to facilitate the wide dissemination of tools and results generated by the Centers' research programs. To support this mandate MAGNet has developed the genomics Workbench, geWorkbench (http://www.geworkbench.org/), a freely available Java application that provides access to an integrated suite of genomics tools produced by MAGNet investigators as well as by external contributors. It is developed on top of an open-source, extensible component architecture specifically designed to facilitate the rapid development of new modules and to support the easy integration of pre-existing tools. By providing a framework to integrate the various MAGNet tools and databases, geWorkbench serves as the main vehicle for disseminating the Center's scientific and technological production to the research community.

At present, geWorkbench integrates over 50 individual components, covering a wide range of genomics domains. For microarray gene expression analysis, several major file formats and chip types are supported. Many filtering and normalization options are available and there are links to several annotation sources, including Affymetrix annotations, caBIO pathways and Gene Ontology terms. Also available are algorithms for differential expression analysis, hierarchical clustering, self-organizing maps, class prediction, regulatory network reconstruction, etc. Sequence support includes BLAST, pattern discovery, transcription factor mapping, and syntenic region analysis. A wide variety of visualizations modules accompany these tools. Additionally, components to support protein structure visualization and analysis are under active development, leveraging one of the major scientific strengths of the Center.

geWorkbench utilizes standards-based middleware grid technologies (such as those developed by the caBIG initiative, https://cabig.nci.nih.gov/, among others) to provide seamless access to remote data, annotation and computational resources thus enabling researchers with limited local resources to benefit from available public infrastructure which otherwise would have been out of their reach or/and would have required a non-trivial level of technical know-how in order to utilize.



geWorkbench: Using a component architecture it allows individually developed plug-ins to be configured into complex bioinformatic applications.

BIOPHYSICAL MODELING OF GENE REGULATORY NETWORKS WITH MATRIXREDUCE

HARMEN BUSSEMAKER LAB

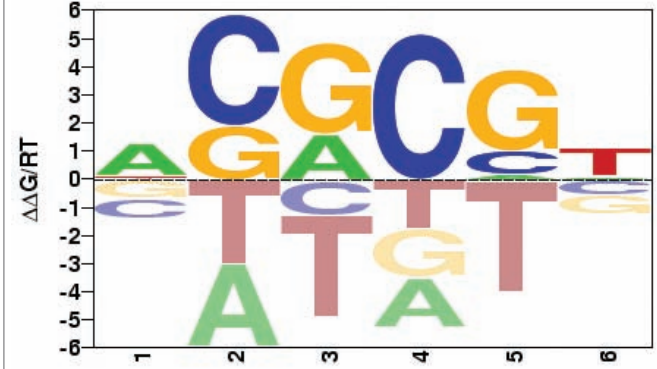
Many algorithms exist for finding sequence "motifs" in nucleotide sequences. The most accurate representation of sequence specificity takes the form of a position-weight matrix (PWM) or position-specific scoring matrix (PSSM). Such matrices model the sequence variability in a collection of aligned binding sites for a particular transcription factor. Algorithms for discovering weight matrices usually require the user to make ad hoc parameter choices, such as how to delineate the set of sequences that will be searched, how to define the statistical properties of "random" nucleotide sequences, and how to pick a threshold for the weight matrix score when predicting binding sites. Barrett Foat, a graduate student in the Department of Biological Sciences, and Harmen Bussemaker, one of the faculty members of C2B2/MAGNet have developed a weight matrix discovery method that avoids these complications. Their algorithm, named MatrixREDUCE, uses a biophysical model for protein-nucleotide interaction that predicts probe binding affinities from the non-coding sequence associated with each probe. It represents sequence specificity in the form of a position-specific affinity matrix (PSAM), whose parameters, determined by fitting the model to a single genomewide mRNA expression profiling or "ChIP-chip" experiment, correspond directly to differences in binding free energy. MatrixREDUCE uses the data for all probes – not just a subset – and no "background" frequencies need to be defined. The inferred PSAM can be used to convert any nucleotide sequence to a single base-pair resolution (relative) binding affinity profile.

[1] H.J. Bussemaker, B.C. Foat, and L.D. Ward (2007). Predicting genomewide mRNA expression: From Modules to molecules. Annual Reviews in Biophysics and Biomolecular Structure.

[2] B.C. Foat, A.V. Morozov, and H.J. Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE yields experimentally verified relative binding affinities (2006). Bioinformatics 22(14): e141-9 (Proceedings of ISMB 2006 conference).

The MatrixREDUCE software can be downloaded from:

http://bussemakerlab.org/software/MatrixREDUCE/



Visualizing with MatrixREDUCE: Position-specific affinity matrix inferred by MatrixREDUCE, represented as an "affinity logo" in which the height of the letters corresponds directly to differences in binding energy

CHROMOSOME EVOLUTION

KENNETH ROSS LAB

Why do some groups of species have widely varying karyotypic features (such as the number of chromosomes) while other related groups have a relatively conserved karyotype? We have proposed a novel hypothesis: At least part of the variation is caused by a species' exposure to alpha radiation in its natural environment. Most natural alpha radiation comes from decay progeny of radon. Exposure is particularly high below ground, and is also elevated on plant surfaces due to deposition by rain.

A survey of karyotypic variation in nature provides support to this hypothesis. Burrowing animals (such as gophers, rabbits, burrowing birds, foxes) have a widely varying karyotype, while their surface-resident relatives (tree squirrels, hares, non-burrowing birds, wolves) have a conservative karyotype. Herbivores have higher karyotpic variation than carnivores, with some interesting exceptions. For example, camels have a conserved karyotype, consistent with the hypothesis since they inhabit regions with low rainfall. Previously unexplained observations, such as that mole-rat taxa show elevated rates of chromosomal speciation in seismic fault zones, can also be explained since radon emissions are known to be elevated in sheared fault zones.

[1] K. A. Ross, "Alpha Radiation is a Major Germ-Line Mutagen over Evolutionary Timescales," Evolutionary Ecology Research, 8(6), 2006, pages 1013-1028.

Columbia University
Herbert Irving Cancer Research Center
1130 St. Nicholas Avenue
New York, NY, 10032

Winter 2008
Issue No. 1, Vol. 1